Development and Deployment of Research Data Preservation Policy at a Japanese Research University in 2016

AUTHOR(S):

Aoki, Takaaki; Kajita, Shoji; Akasaka, Hirokazu; Takeda, Hagane

CITATION:

Aoki, Takaaki ...[et al]. Development and Deployment of Research Data Preservation Policy at a Japanese Research University in 2016. 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) 2017: 120-123

ISSUE DATE:

2017

URL:

http://hdl.handle.net/2433/230692

Development and Deployment of Research Data Preservation Policy at a Japanese Research University in 2016

Takaaki AOKI, Shoji KAJITA
Institute of Information Management and Communication
Kyoto University
Sakyo, Kyoto, 6068501, JAPAN

Hirokazu AKASAKA, Hagane TAKEDA
Planning and Information Management Department
Kyoto University
Sakyo, Kyoto, 6068501, JAPAN

***Abstract*—This paper reviews the policy development and deployment for research data preservation at Kyoto University, Japan, during fiscal year (FY) 2016. The university's regulations and guidelines for research integrity and data preservation were formulated in FY2014 and 2015, in response to statements from the Ministry of Education, Culture, Sports, Science and Technology, and supplemental comments by Science Council of Japan (in FY2014). In FY2016, several departments at KU developed a prototype system for research data preservation using open source web frameworks with preference to agility of deployment rather than robustness. The prototype system also worked as the proof of concept for a full-service research data preservation system, which is due to be deployed by KU's central IT department in FY2017 for university-wide use.**

Keywords—Research Data Management, Data Preservation, Research Integrity, Archive System

I. INTRODUCTION

In 2013 and 2014, Japanese academia was shocked by high-profile incidents of scientific misconduct. Members of the academy at research institutes and their colleagues at government offices and academic societies were called upon to help with the urgent reconstruction and development of policy, guidelines, and procedures to ensure research integrity. In particular, a mandate for preservation of research data was issued for both researchers and research institutes. In this paper, we briefly review the actions to ensure research integrity taken at Kyoto University (KU), one of the largest national research universities in Japan. We especially focus on the policy development and deployment of research data preservation, and the development of a prototype archiving system and the full-service research data preservation system expected to be developed from it.

II. POLICY DEVELOPMENT FOR RESEARDH DATA PRESERVATION

A. Actions by the Japanese Government

In 2014, the Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan revised the old previous guideline against misconduct and issued "Guidelines for Responding to Misconduct in Research" [1]. The new guideline mandates on that research universities and institutes to formulate a develop regulation requirements that researchers should preserve research data for a certain period and disclose according totheir research data upon request. The Science Council of Japan organized a working group to discuss practical action items along the MEXT guideline. In At the end of fiscal year 2014, the working group issued a supplemental comment on the MEXT guideline, "Enhancing the Integrity of Scientific Research (Response)" [2], which suggests guidelines for the classification of the research data and decides proposing the a preservation term of more than 10 years for general research data.

B. Actions at Kyoto University

Kyoto University is one of Japan's largest national research universities, consisting of more than 2,700 academic staff and 9,000 graduate students. KU promptly responded to the MEXT guideline and formulated the regulation "Promoting Research Integrity Regulations of Kyoto University" [3] in FY2014. This regulation mainly states the university's obligation to ensure, and role in ensuring, research integrity as described in MEXT's guideline.

In FY2015, the detailed roles and obligations for research data management were issued as "Matters ruled for the Preservation and Disclosure of Research Data as defined in Article 7- 2 of the Regulations regarding Promoting Research Integrity of Kyoto University" [4], which also refers to the supplemental comments by the Science Council of Japan. This

regulation demands that each department formulate procedures for research data preservation.

III. Development of A prototype research data archive system (FY2016)

A. *Cases of KU research departments and central IT division*

For many researchers, it is a natural assumption that their research data should be kept for as long as possible, protected from data loss and corruption due to any accidental or artificial reason. However, this becomes extremely difficult to achieve when data preservation is mandated to every researcher. Ensuring the availability and integrity of research data for more than 10 years goes beyond an individual researcher's personal IT skills. According to the above regulation for data preservation issued by KU, the initial responsibility for data preservation is on the individual researcher, while the department is obliged to audit and support the data preservation process by its researchers. Thus, during FY2015 and 2016, some departments promptly started to develop data preservation systems.

For example, KU's Graduate School of Engineering decided to introduce a prototype research data archiving system with a minimum set of requirements. Initially, to speed up development time, the prototype was designed to secure only native digital data related with (a) published papers with a digital object identifier (DOI), and (b) submitted doctoral dissertations.. The archive system was constructed as an extension of the Plone web content management system [5], as Plone had been adopted as the standard CMS for the school and had suitable features for research data archiving, such as tagging metadata and sharing archived data among local users.

The Institute of Information Management and Communication (IIMC), developed as KU's central IT service division, also developed a prototype for a research data archiving system in collaboration with the Graduate School of Engineering, using Ruby on Rails [6] technology. This IIMC prototype was not only designed for university-wide use but also to serve as proof of concept toward a full-service data archiving system to be developed later.

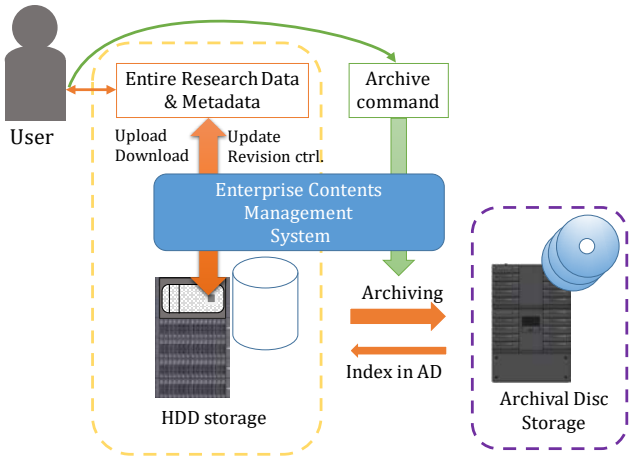

Fig. 1. Schematic Concept of data preservation system using ECM and archive storage.

B. *Issues on on prototype data archiving systems and services*

Both systems launched around April 2016. However, as of March 2017, the amount of data archived in these systems is extremely low considering the actual research activities by relevant faculty and researchers. In the Graduate School of Engineering system, , 57 articles and 14 GB of research data has been archived by 27 unique users, though the total relevant research faculty numbers 430. The IIMC system holds 154 articles and 34 GB submitted by 33 unique users, also a very low ratio in relation to KU's total academic faculty.

The reason for low usage of these prototype archiving systems is due to weaknesses in both systems' architecture system operation. In other words, neither organization can properly support the stability of its archiving system. In general, data preservation entails serious issues in deciding how/what to guarantee, or exempt, between the system provider and users. Agreement between providers and users may require a much larger amount of evidence, and more time to develop.

Cost is another problem for data preservation. From the context of research integrity, the archived data is kept un-published and unused for more than 10 years, which implies data preservation is only the cost, and no benefit is to be expected. System providers as well as institute administrations must consider how to build and operate the archiving system at low cost.

IV. Archiving system utilizing enterprise content management system and optical disc archiver

A. *System overview*

In addition to operating the prototype system, the IIMC designed a stable and cost-effective research data archiving system in FY2016. This latter system consists of an enterprise content management (ECM) system and an optical disc storage system. The schematic concept is shown in Fig. 1. The term ECM is widely defined as in [7] and [8]; 'The technologies used to capture manage, store, deliver and preserve information to support business processes.' From the viewpoint of research data preservation, the following functions are required by ECM, sometimes called 'document management';

- access control and auditing
- metadata tagging
- revision management
- searchable content and metadata

However, it is difficult to secure research data for the long term with ECM due to the shorter lifetime of ECM system hardware, software and database structure compared to the required time for preservation. This shorter lifetime entails that integrity testing may occur more frequently with each update of the system hardware, database format or ECM software.

This problem could be solved by connecting ECM with other long term preservation archiving systems in which retrieved data is archived on classical and open data formats and file systems.

For this system, the IIMC utilizes an Oracle WebCenter Content (OWCC) [9] and FUJITSU Eternus DA700 data archiver [10] . OWCC is an instance of ECM software which provides the requisite functions mentioned above. The DA700 is a disc array system consisting of an Archival Disc. The Archival Disc is 'write once read many (WORM)' media and guarantees more than 50 years of data preservation time. Moreover, discs are assembled in a cartridge and may incorporate RAID5 or 6 to improve redundancy.

B. Data preservation procedure

The typical scenario for data operation and archiving is as follows.

1. Users can create folders and upload their research data on OWCC. Users may organize their research data using OWCC functionality, such as tagging metadata, utilizing revision control, or sharing collaborators for local use.
2. A user can issue the 'archive' command on any folder under his/her administration. The archive command retrieves all content within a given folder and its descendants. These contents are copied to DA700 with additional information such as metadata, an access control list, etc. Typically, the archive command is processed as a batch, so a user can cancel it before actual operation. If content has several revisions on OWCC, the content's owner can choose copying the latest version only or all revisions to DA700.
3. When the data copy from OWCC to DA700 is finished, index information on DA700 is included with the source content as metadata. Additionally, the access to the source content on OWCC is set to 'read only,' including for the owner of the content. This process ensures the contents on OWCC and DA700 are the same. The content owner may retrieve write or administrative access control with several steps on OWCC, and make a copy on DA700 again. This feature enables the user to keep archive revisions on DA700, as well as reducing frequent copying to DA700.

Under the system operation policy, no user can access the data archived in DA700. This means that the data in DA700 is treated as a 'dark archive' and also ensures fairness in the research data preservation procedure.

C. Metadata for research data preservation in ECM

On OWCC, all content is categorized by a 'document type' property, and users can attach any metadata fields to the content regardless of its document type. For research data preservation, metadata are implemented for the purpose of 'linkage to published material' or 'authorship and location of responsibility'. In order to distinguish related published materials, the following attributes are implemented.

- 'archive title'
- 'authors'
- 'published media'
- 'DOI or other identifiers'
- 'date of publication'
- 'accuracy of date of publication'

The 'accuracy of date of publication' is chosen from the following options:

- year, month and date are correct
- year and month are correct, but date is not fixed
- year is correct, month is between April and December
- year is correct, month is between January and March
- year is correct, no accurate month nor date information

The 3rd and 4th options are designed to identify the fiscal year in Japan.

Additionally, the some metadata fields are included to ensure authorship and responsibility.

- 'system_login_id'
- 'system_login_name'
- 'system_login_organization'

These fields are automatically populated from the university directory server as determined by the login user on the OWCC system. However, there are many cases in which different authorship and organization information may be required. For instance, when data is uploaded by a laboratory staff on behalf of the PI. Another example would be when a researcher belongs to multiple departments, but the uploaded research data should be assigned to one responsible department over the others. Therefore, additional metadata are also available and can be modified by contents owners, including:

- 'editor_id'
- 'editor_name'
- 'editor_organization'

SUMMARY

KU's ECM based data preservation system will launch in FY2017 for university-wide use. The system is currently in preliminary use to test and discuss reliability, usability, performance and operation costs. Additionally, the procedure for accessing the archival disc storage needs development. The contents of the archival storage disc are treated as a dark archive, so that several steps for authentication and authorization are required. Though the system is being developed by KU's central IT division, users should follow the different data preservation procedure adopted by their own department. Further, the system's specifications and operation rules should be inspected and qualified by each department.

Issues with reliability and cost still remain. The best method to solve these issues is simply to increase the overall use of the

system. Increase of system use, as well as exposure to more users, may result in the discovery of previously undetected, small errors and the improvement of system usability. However, the system is designed to preserve data only for the purpose of verifying research integrity, it is closed to researchers and rarely accessed. The archived contents are only accessed when there is a question of research misconduct. Going forward, it is important that the institution discusses how to extend and apply this system for data publishing, oriented to the concept of Open Access, Open Data and Open Science.

ACKNOWLEDGMENT

The authors thank Prof. Takaaki Komura, Institute for Information Management and Communication and Mr. Takahiro Okunaka, Center for Information Technology, Graduate School of Engineering, Kyoto University, for details in the design and status of each department's research data archiving system.

REFERENCES

[1] Ministry of Education, Culture, Sports, Science and Technology G. Eason, "'Guidelines for Responding to Misconduct in Research' (Adopted August 26, 2014 by Ministry of Education, Culture, Sports, Science and Technology (MEXT))", 2014 [online], Available: http://www.mext.go.jp/a_menu/jinzai/fusei/1359618.htm [10-Apr-2017]

[2] Sicence Council of Japan, "科学研究における健全性の向上について(回答) [Enhancing the Integrity of Scientific Research (Response)]", 2015 [online], Available: http://www.scj.go.jp/ja/member/iinkai/kenzensei/pdf/kenzensei-kaito.pdf [10-Apr-2017]

[3] Kyoto University, "Promoting Research Integrity Regulations of Kyoto University", 2015 [online], Available http://www.kyoto-u.ac.jp/en/research/ethic/research_guide/documents/research-integrity-regulations201503.pdf [10-Apr-2017]

[4] Kyoto University, "Matters ruled for the Preservation and Disclosure of Research Data as defined in Article 7- 2 of the Regulations regarding Promoting Research Integrity of Kyoto University", 2015 [online]. Available: http://www.kyoto-u.ac.jp/en/research/ethic/research_guide/documents/research_data_en150730.pdf [10-Apr-2017]

[5] Plone Foundation, "Plone CMS: Open Source Content Management — Site" Internet: www.plone.org [10-Apr-2017].

[6] "Ruby on Rails - A web-application framework that includes everything needed to create database-backed web applications according to the Model-View-Controller (MVC) pattern.", Internet: http://rubyonrails.org/ [10-Apr-2017]

[7] Association for Information and Image Management, "What is Enterprise Content Management (ECM)?" Internet: http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx# [10-Apr-2017]

[8] U. Kampffmeyer, "ECM Enterprise Content Management", 2006 [online]. Available: http://www.project-consult.de/Files/ECM_White%20Paper_kff_2006.pdf [10-Apr-2017]

[9] Oracle, "Oracle WebCenter Content", Internet: www.oracle.com/technetwork/middleware/webcenter/content/overview/index.html 10-Jan-2013 [10-Apr-2017]

[10] Fujitsu limited "FUJITSU Storage ETERNUS DA700 Data Archiver", Inernet: http://www.fujitsu.com/jp/products/computing/storage/data-protection/da700/ [10-Apr-2017].